EXHIBIT AA

BIOMETRICS 45, 1309–1322
December 1989

READER REACTIONS

Confounding in Epidemiologic Studies*

Sander Greenland (Department of Epidemiology, UCLA School of Public Health, Los Angeles, California 90024-1772, U.S.A.)

I would like to point out that there is a large body of literature, not referenced by Wickramaratne and Holford (1987, *Biometrics* **43,** 751–765), concerned with the adequacy of control groups and measures of confounding. I am aware of three somewhat nonoverlapping branches of this literature.

The first branch, exemplified by Rubin (1974, 1977), Rosenbaum and Rubin (1983, 1985), Rosenbaum (1984, 1987), and Holland (1986), comprises the extensive work of these authors on problems of defining, detecting, preventing, and adjusting for inadequacies of the control group. Such inadequacies are formalized under the concept of "nonignorability of treatment assignment" (Rosenbaum, 1984). I suspect that Rubin's formalization is the best currently available for dealing with problems of causal inference, such as confounding. My only serious objection to this literature is that at some points it proposes to check for confounding by means of significance tests [see, for example, the test of strong ignorability in Rosenbaum (1984, §5)]. In nonrandomized studies, it is the hypothesis that confounding exceeds a specified level (not the hypothesis of nonconfounding) that must be rejected before one can confidently proceed with inference regarding treatment effects. Thus, if one insists on doing a frequentist test for confounding, the logical choice is an *equivalence* test rather than a significance test (Greenland, 1989). Admittedly, an equivalence test requires one to parameterize the degree of confounding, but I view this complication as a benefit of equivalence testing: Given that some confounding is almost always present in nonrandomized studies, one should test whether the amount present is worth worrying about, rather than test a certainly false null hypothesis.

A second branch, exemplified by Greenland and Robins (1986) and Robins and Morgenstern (1987), developed from the Miettinen and Cook (1981) article discussed by Wickramaratne and Holford. These authors examine the relation of formal concepts such as exchangeability (comparability) and collapsibility to intuitive notions of confounding. Like Wickramaratne and Holford, these authors point out the need to distinguish the phenomena of comparability and collapsibility when attempting to deal with confounding in a formal mathematical framework. Greenland and Robins (1986) also show that confounding and comparability can be defined without any reference to covariates, a point not mentioned by Wickramaratne and Holford, but which follows directly from Rubin's formalization.

A third branch, exemplified by Gail (1986, 1988) and Chastang, Byar, and Piantadosi (1988), concerns adjustment for balanced covariates. In particular, Gail characterizes the class of models under which balance on a covariate (comparability with respect to the covariate) implies that the point estimator is collapsible over the covariate (here, "collapsible" means the unadjusted estimator is asymptotically unbiased for the parameter of interest). He also characterizes a subclass of this class in which failure to adjust for the covariate leads to invalid hypothesis tests and confidence intervals. The models in this

* *Editor's note*: This is a collection of three reactions to a paper published in this journal in 1987, together with the original authors' response.

subclass demonstrate that balance and point-estimate collapsibility on a covariate are not always jointly sufficient conditions for ignoring the covariate (although they will often be approximately so). My chief objection to this literature is its tendency to identify effects with regression model coefficients; this identification results in model dependence of causal concepts such as "effect" and "confounder." For several reasons, I regard model dependence of such concepts as undesirable and unnecessary: (1) In epidemiology, at least, the true regression model is never known; (2) causal concepts can be defined nonparametrically, as in the other two literature branches discussed above; and (3) an *estimate* of effect may be derived using a parametric model even if the effect itself is *defined* nonparametrically.

References

Chastang, C., Byar, D., and Piantadosi, S. (1988). A quantitative study of the bias in estimating the treatment effect caused by omitting a balanced covariate in survival models. *Statistics in Medicine* **7,** 1243–1255.

Gail, M. H. (1986). Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In *Modern Statistical Methods in Chronic Disease Epidemiology*, S. H. Moolgavkar and R. L. Prentice (eds). New York: Wiley.

Gail, M. H. (1988). The effect of pooling across strata in perfectly balanced studies. *Biometrics* **44,** 151–162.

Greenland, S. (1989). Comment: Cautions in the use of preliminary-test estimators. *Statistics in Medicine* **8,** 669–673.

Greenland, S. and Robins, J. M. (1986). Identifiability, exchangeability, and epidemiologic confounding. *International Journal of Epidemiology* **15,** 413–419.

Holland, P. W. (1986). Statistics and causal inference (with Discussion). *Journal of the American Statistical Association* **81,** 945–970.

Miettinen, O. S. and Cook, E. F. (1981). Confounding: Essence and detection. *American Journal of Epidemiology* **114,** 593–603.

Robins, J. M. and Morgenstern, H. (1987). The foundations of confounding in epidemiology. *Computers and Mathematics with Applications* **14,** 869–916.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of strongly ignorable treatment assignment. *Journal of the American Statistical Association* **79,** 41–48.

Rosenbaum, P. R. (1987). The role of a second control group in an observational study (with Discussion). *Statistical Science* **2,** 292–316.

Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B* **45,** 215–218.

Rosenbaum, P. R. and Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics* **41,** 103–116.

Rubin, D. B. (1974). Estimating the causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology* **66,** 688–701.

Rubin, D. B. (1977). Assignment of treatment group on the basis of a covariate. *Journal of Educational Statistics* **2,** 1–26.

Wickramaratne, P. J. and Holford, T. R. (1987). Confounding in epidemiologic studies: The adequacy of the control group as a measure of confounding. *Biometrics* **43,** 751–765.

Received February 1989; *revised May and July* 1989.

Paul W. Holland (Educational Testing Service, Princeton, New Jersey 08541-0001, U.S.A.)

Wickramaratne and Holford (1987, *Biometrics* **43,** 751–765; herein W&H) present a mathematical model for confounding in epidemiologic studies that is based on the notion of "the adequacy of the control group" rather than on the criterion of the collapsibility in multiway contingency tables. I agree with them that collapsibility is not the fundamental idea underlying confounding, and their intuitive, *informal* description of confounding agrees in many respects with my own views. However, I think that the *formal model* for confounding in the population they present is incomplete and unable to express important features of the problem that they *do* express informally in their discussion. Their model is based on a hypothetical population of individuals who are exactly like the exposed population but who are not exposed. In my opinion, this general approach can deal with only a few of the complexities of causal inference in epidemiologic studies and eventually fails to express all of the relevant ideas. I advocate the use of a framework originally developed for the analysis of experiments (Neyman, 1935) and extended to apply to observational studies (Rubin, 1974, 1978) that I call "Rubin's model" (Holland, 1986). The use of a new notation should not be advocated for its own sake or for purely aesthetic reasons. Hence, my goal here is to show that Rubin's model can clarify both W&H's notation and their discussion of confounding.

1. Rubin's Model

In order to avoid confusion between W&H's notation and mine, I will use different symbols for the various variables and will then translate between the two systems insofar as this is possible.

A population of individuals underlies both analyses. I will be explicit and call this population U, denoting a generic individual in U by $u \in U$. For W&H, U is the *study* population. A variable S is defined on each $u \in U$ so that $S(u) = e$ if u is *exposed* to the causal agent of interest and $S(u) = c$ if u is not so exposed. When I need to refer to an unspecified exposure condition I will use lowercase s to denote either e or c. So far this notation is similar to that of W&H. Now comes the major difference. The problem is to create a notation that can describe the disease status of each individual in U. The issue is that each individual really has two potential disease statuses—one that will occur if the individual is exposed to the causal agent of interest and another that will occur if the individual is not so exposed. Thus, disease status is really a function of (i.e., depends on) two quantities: an individual, u, and an exposure condition, s. This leads us to define the disease status, $Y(u, s)$, for all pairs (u, s) with $u \in U$ and $s = e$ or c by

$$Y(u, s) = \begin{cases} 1 & \text{if, were } u \text{ to be exposed to condition } s, u \text{ would develop the disease and,} \\ 0 & \text{if, were } u \text{ to be exposed to condition } s, u \text{ would not develop the disease.} \end{cases}$$

It is notationally a bit simpler to put s into a subscript as follows:

$$Y_s(u) = Y(u, s),$$

in which case we may talk about the two "potential versions" of Y, Y_e and Y_c.

One might object that in any real epidemiologic study one does not *observe* two versions of Y. Instead, a single disease status is obtained for each individual. Upon reflection this is seen to be a superficial view. For individuals exposed to e, Y_e is observed and Y_c is not, but for those exposed to c, Y_c is observed and Y_e is not. Thus, the fact that only one version of Y is *observed* is quite unrelated to the fact that two possibly different values of Y *could have been observed.* I denote the *observed disease status* of each individual by Y_S and the exposure condition of each individual by S. Hence, when $S = e$, we have $Y_S = Y_e$ and when $S = c$, we have $Y_S = Y_c$. Thus, this notation can express what is observed, i.e., (Y_S, S), but it can also express the causal parameters that are behind the observed data. These causal parameters are the individual-level causal effects defined as the difference

$$Y_e(u) - Y_c(u).$$

Note that $Y_e(u) - Y_c(u) = 1$ only when u will get the disease if exposed to the causal agent but will not otherwise. Also, $Y_e(u) - Y_c(u) = 0$ if the disease status of u is the same regardless of the exposure condition. Finally $Y_e(u) - Y_c(u) = -1$ if u gets the disease if u is not exposed but does not get the disease if exposed—a situation that may sometimes be ruled out based on subject-matter knowledge.

Multiple versions of the response variable, Y_e and Y_c, have been used in discussions of randomization distributions in randomized experiments since Neyman (1935). One of Rubin's contributions was to show their importance to *all* discussions of causal inference including nonexperimental studies.

Even though Rubin's model can express the fundamental individual-level causal effects, in applications of the model we usually aggregate these into the average causal effect or ACE, as follows (Holland, 1988). If $\mathrm{E}(\cdot)$ denotes expectation or population average (over U) then the ACE is defined as

$$\begin{aligned}\mathrm{ACE} &= \mathrm{E}(Y_e - Y_c)\\ &= P(Y_e = 1) - P(Y_c = 1),\end{aligned} \tag{1}$$

where $P(\cdot)$ denotes proportions of individuals in the study population.

In epidemiologic analyses it is customary to restrict the average causal effects to those individuals who were exposed, the so-called "exposed population." This may be denoted in the following obvious way in Rubin's model using conditional expectations and conditional probabilities:

$$\begin{aligned}\mathrm{ACE}(S = e) &= \mathrm{E}(Y_e - Y_c \mid S = e)\\ &= P(Y_e = 1 \mid S = e) - P(Y_c = 1 \mid S = e).\end{aligned} \tag{2}$$

At this point it may be helpful to connect this development to that of W&H. They use E and $\bar{E}$ to refer to the subpopulations of individuals for which $S(u) = e$ and $S(u) = c$, respectively. They use D and $\bar{D}$ to refer to the subpopulations of individuals whose observed disease statuses are $Y_S = 1$ and $Y_S = 0$, respectively. They denote the disease rates among the exposed and unexposed individuals by $P(D \mid E)$ and $P(D \mid \bar{E})$, whereas in Rubin's model these rates are, respectively,

$$P(Y_S = 1 \mid S = e) = P(Y_e = 1 \mid S = e)$$

and

$$P(Y_S = 1 \mid S = c) = P(Y_c = 1 \mid S = c).$$

Instead of having two versions of Y, Y_e and Y_c, as in Rubin's model, W&H attempt to accomplish the same end by introducing the quantity $P'(D \mid \bar{E})$, which they define as the "hypothetical proportion of the exposed individuals who would have developed the disease even if they had not been exposed." It is not clear in their model just what kind of mathematical object P' is supposed to be. Is P' a probability like P defined as proportions of individuals in U? If so, then why not use P? It is clear that P' has the wrong thing in the conditioning (i.e., to the right of the vertical bar) because $\bar{E}$ does not describe the subgroup of *exposed* individuals (to which the definition of P' refers). Rubin's model helps clarify what $P'(D \mid \bar{E})$ is: $P'(D \mid \bar{E})$ is just $P(Y_c = 1 \mid S = e)$, i.e., the proportion of the exposed individuals who would have developed the disease had they not been exposed. What is nice about Rubin's model is that it is clear from the notation that $P(Y_c = 1 \mid S = e)$ is not a directly observable quantity because it requires information about what would have happened to the exposed individuals had they not been exposed to the causal agent of interest. The probability $P(Y_c = 1 \mid S = e)$ is "counterfactual" in the sense that the event in the conditioning ($S = e$) precludes us from ever obtaining direct information about the event whose probability is sought ($Y_c = 1$). Issues of counterfactuality pervade many discussions of causal inference (Glymour, 1986).

The *conditional* average causal effect, ACE($S = e$), defined in (2), is exactly what W&H call the risk difference:

$$\begin{aligned} RD = \Delta &= P(D \mid E) - P'(D \mid \bar{E}) \\ &= P(Y_e = 1 \mid S = e) - P(Y_c = 1 \mid S = e) = \text{ACE}(S = e). \end{aligned} \tag{3}$$

Thus, Rubin's model clarifies the nature of Δ as an average causal effect over the subpopulation of the exposed individuals. This is not at all evident from W&H's notation, even though it is clearly expressed in their informal description.

The ACE($S = e$) is the causal parameter of interest, but we can never calculate it directly because of the counterfactual nature of $P(Y_c = 1 \mid S = e)$ and $P'(D \mid \bar{E})$. Instead, we compute the *prima facie average causal effect* or FACE defined as

$$\begin{aligned} \text{FACE} &= \text{E}(Y_S \mid S = e) - \text{E}(Y_S \mid S = c) \\ &= P(Y_e = 1 \mid S = e) - P(Y_c = 1 \mid S = c). \end{aligned} \tag{4}$$

In W&H's notation,

$$\text{FACE} = P(D \mid E) - P(D \mid \bar{E}),$$

the *observed* risk difference. For both W&H's model and Rubin's model the basic question is the same: Is the FACE (which can be estimated from the data in a cohort study) equal to the ACE($S = e$) (which is the causal parameter of interest)? W&H put this in terms of substituting $P(D \mid \bar{E})$ for $P'(D \mid \bar{E})$ in Δ and observe that the substitution is correct if and only if

$$P(D \mid \bar{E}) = P'(D \mid \bar{E}), \tag{5}$$

which, in Rubin's model, is expressed as

$$P(Y_c = 1 \mid S = c) = P(Y_c = 1 \mid S = e). \tag{6}$$

W&H define confounding to exist in the population if (6) fails to hold. In terms of Rubin's model, confounding exists in the population if the ACE($S = e$) and the FACE are unequal. Combining (3) and (4), this is equivalent to condition (6) failing, i.e., that over the population U, the variable Y_c is not statistically independent of S. In a large, randomized prospective study, S is constructed to be independent of Y_c (and of Y_e, too) and there is no confounding in the population (Holland and Rubin, 1988).

2. The Role of Other Variables

In discussions of the need to adjust analyses for "potential confounders" it is necessary to introduce other variables into the picture. W&H do this by letting C_i denote the ith category defined by a discrete variable. I will let X be the discrete variable with possible values denoted by i. However, even here careful attention must be given to the notation. In principle, whenever a new variable is introduced into Rubin's model it is necessary to treat it like Y was, i.e., X is defined on the pairs, (u, s), of individuals and potential exposure conditions. $X(u, s)$ is the value of X that would be measured on u if u were exposed to condition s. Again a subscript notation is helpful, i.e., we define $X_s(u)$ by

$$X_s(u) = X(u, s). \tag{7}$$

However, a new issue can arise due to the possible relationship of X to the exposure condition, s. When $X_s(u)$ does *not* depend on s we drop s from the notation and call $X(u)$ a *covariate*. Covariates are variables that are not affected by exposure to the causal agent of interest. Common medical examples of covariates are age, sex, and variables measured prior to the exposure of u to e or c. Variables measured after exposure to e or c are *always potentially affected by this exposure* and their dependence on s must always be carefully considered. W&H discuss covariates but since their notation cannot express the dependence of C on s, their discussion remains informal—i.e., "This assumption implies that for each individual in the exposed population the value of C is unaffected by exposure; hence, C cannot be an intervening variable in the causal pathway between exposure and disease" (p. 753). What they need to be able to express in their model, and cannot, is that C is a covariate in the sense defined above for X_s. But it is not necessary, initially, to assume that C is a covariate and I will not assume it is until later.

The basic difference between Rubin's model and W&H's approach can be expressed as follows. W&H use subpopulations of exposed individuals and unexposed individuals and a hypothetical population that is just like that of the exposed individuals but which is unexposed to the causal agent of interest. Rubin's model, on the other hand, uses one population and two different potential measurements, Y_e and Y_c, that might be made on each individual in this population. Only one of these potential measurements, Y_S, can be observed for each individual but the effect of the causal agent on each individual is defined as the difference between Y_e and Y_c. In Rubin's model the effect of a causal agent is defined at the level of the individuals and then aggregated by averaging to obtain the ACE parameter that describes the overall effect of exposure to the causal agent on the population of individuals of interest. The result, ACE($S = e$), is exactly the same causal parameter defined by W&H as the difference between the disease rate in the exposed population and the disease rate in the hypothetical "unexposed" exposed population.

The mathematical structure of Rubin's model is completely transparent since it is based, at bottom, on the joint distribution of the five-dimensional vector (S, Y_e, Y_c, X_e, X_c) over the population U. In contrast, W&H need to define a new probability apparatus, $P'(\cdot)$, which looks like a conditional probability with the wrong thing in the conditioning and which does not have any clear-cut rules of combination with the more straightforward P probabilities. In my opinion causal inference is complicated enough without the introduction of a new, nonstandard, and undefined mathematical notation. Rubin's model can express all the ideas we need for simple cohort studies and can be expanded to accommodate case–control studies (Holland and Rubin, 1988) and indirect causation (Holland, 1988). Robins (1987) gives a model that involves time-varying covariates and which is clearly in the spirit of what I call Rubin's model.

3. A Comparison of W&H and Rubin's Models

W&H first make the following very important assumption about C:

$$P(D \mid \bar{E}C_i) = P'(D \mid \bar{E}C_i) \quad \text{for all } i. \tag{8}$$

The definition of the left-hand side of (8) is straightforward, but that of the right-hand side is more complicated because it involves P'. W&H describe $P'(D \mid \bar{E}C_i)$ as ". . . the hypothetical probability of disease of an exposed individual, given that this individual is in stratum C_i of the exposed population, if the individual had not been exposed." This could mean two different things in terms of Rubin's model. On the one hand, it might mean

$$P'(D \mid \bar{E}C_i) = P(Y_c = 1 \mid S = e, X_e = i) \tag{9}$$

or, on the other hand, it might mean

$$P'(D \mid \bar{E}C_i) = P(Y_c = 1 \mid S = e, X_c = i). \tag{10}$$

Interpretation (9) emphasizes being "in stratum C_i of the exposed population" while (10) emphasizes the counterfactual nature of "if the individual had not been exposed." In an earlier version of this note, I used (9) as the interpretation, but Wickramaratne and Holford (personal communication) have pointed out to me that (10) is what they had intended. Using (10) to define $P'(D \mid \bar{E}C_i)$, we can express (8) in terms of Rubin's model as

$$P(Y_c = 1 \mid S = c, X_c = i) = P(Y_c = 1 \mid S = e, X_c = i) \quad \text{for all } i, \tag{11}$$

where I have not yet assumed that X_s is a covariate. Assumption (11) is not discussed by W&H but it is a very important assumption and underlies all of their analyses. Equation (11) says that Y_c and S are conditionally independent over U given X_c. This is related to the condition of strong ignorability defined by Rosenbaum and Rubin (1983), and it is probably the most important type of assumption that is made in all discussions of causal inference in nonrandomized studies (Rubin, 1974).

W&H then give two conditions, each of which ensures that there is no confounding in the population in the sense of equation (6). These are, in their notation:

(i) $P(D \mid \bar{E}C_i) = P(D \mid \bar{E})$ for all i; and
(ii) $P(C_i \mid \bar{E}) = P'(C_i \mid \bar{E})$ for all i.

In terms of Rubin's model, and not making the assumption that X_s is a covariate, (i) and (ii) may be expressed as

(i′) $P(Y_c = 1 \mid S = c, X_c = i) = P(Y_c = 1 \mid S = c)$ for all i;
(ii′) $P(X_c = i \mid S = c) = P(X_c = i \mid S = e)$ for all i.

Both statements are about (conditional) independence over U. Equation (i′) says that Y_c and X_c are conditionally independent given $S = c$, whereas (ii′) says that X_c and S are independent. Since (i′) involves only Y_c, X_c, and $S = c$, it can be checked in data. However, (ii′) involves the distribution of X_c among those individuals for whom $S = e$ and is therefore not directly testable in the data. Even if we do not assume that C is a covariate, it is easy to show that conditions (i′) and (ii′) are both strong enough in the presence of (11) to imply the equality of $P(D \mid \bar{E})$ and $P'(D \mid \bar{E})$, i.e., no confounding in the population (which is what W&H assert).

The counterfactual nature of (ii′) renders it somewhat impractical since it cannot be directly tested in data. W&H are aware of this and introduce a third condition to make "condition (ii) to be of practical value" (p. 753):

$$P'(C_i \mid \bar{E}) = P(C_i \mid E). \tag{12}$$

The right-hand side of (12) is straightforward, but the left-hand side involves P'. In terms of Rubin's model, (12) is expressed as

$$P(X_c = i \mid S = e) = P(X_e = i \mid S = e). \tag{13}$$

W&H say that (13) "implies that for each individual in the exposed population the value of C is unaffected by exposure." This is not quite correct and is another place where Rubin's model is helpful in clarifying the issues. Equation (13) says that X_c and X_e have the *same distribution* in the exposed population. On the other hand, X is unaffected by exposure if and only if $X_e(u) = X_c(u)$ for all $u \in U$, i.e., if X is a covariate. If X *is* a covariate, *then* (13) must hold so that a better way of wording the relationship might be "if C is unaffected by exposure then (13) holds." It is often easy to know that C is *not* affected by exposure (i.e., is a covariate), in which case (13) will then hold. Equation (13) makes condition (ii) "practical" in the sense that if (13) and (ii) *both* hold then the following *testable* condition will also hold:

$$P(X_S = i \mid S = e) = P(X_S = i \mid S = c) \quad \text{for all } i, \tag{14}$$

where X_S is the observed value of X. Equation (14) is simply the condition that the observed value of X, X_S, has the same distribution in both the exposed and unexposed groups—i.e., X_S is independent of S. Thus, condition (14) can be checked in the data, whereas neither of W&H's conditions (ii) and (13), can be checked because they both involve the counterfactual probability, $P(X_c = i \mid S = e)$. If we reject (14) based on the data, then, in the case when C is a covariate (and therefore (13) must hold), we must also reject W&H's condition (ii) and it becomes testable, i.e., "of practical value." When C is *not* a covariate then (13) may or may not be true but there is no way to directly check it, and W&H's condition (ii) ceases to be directly testable in the data using (14).

4. Final Comment

This note is mainly an attempt to show that the ideas and notation of Rubin's model are based on simple ideas of probability theory, easy to use, and can give useful precision to discussions of causal inference that involve estimating causal effects. Furthermore, it is not W&H's informal discussion of confounding that I would criticize, but rather their use of an unusual and, to me, unnecessary notation to *formalize* it. However, I strongly agree with their conclusion that analyses of contingency tables cannot answer the question of whether a particular variable is a confounder or not. Confounding (i.e., the inequality of the ACE and FACE) is related to collapsibility in the five-way table of (S, Y_e, Y_c, X_e, X_c), which is inherently unobservable, and which is, therefore, not something that a chi-square test can detect. In particular, the important role of untestable assumptions, such as equation (11), cannot be overemphasized. The reader who is interested in more discussions of the use of Rubin's model in epidemiologic analysis is referred to Holland and Rubin (1988).

References

Glymour, C. (1986). Comment: Statistics and metaphysics. *Journal of the American Statistical Association* **81,** 964–966.

Holland, P. W. (1986). Statistics and causal inference (with Discussion). *Journal of the American Statistical Association* **81,** 945–970.

Holland, P. W. (1988). Causal inference, path analysis and recursive structural equations models. In *Sociological Methodology*, C. C. Clogg (ed.), 449–484. Washington, D.C.: American Sociological Association.

Holland, P. W. and Rubin, D. B. (1988). Causal inference in retrospective studies. *Evaluation Review* **13,** 203–231.

Neyman, J. (with Iwaszkiewicz, K. and Kolodziejczyk, S.) (1935). Statistical problems in agricultural experimentation (with Discussion). *Supplement to the Journal of the Royal Statistical Society* **2,** 107–180.

Robins, J. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases* **40,** Supplement, 139S–161S.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66,** 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* **6,** 34–58.

Wickramaratne, P. J. and Holford, T. R. (1987). Confounding in epidemiologic studies: The adequacy of the control group as a measure of confounding. *Biometrics* **43,** 751–765.

Received January 1989; *revised May* 1989.

Nathan Mantel (Department of Mathematics and Statistics, The American University, 4900 Auburn Avenue, Bethesda, Maryland 20814, U.S.A.)

Wickramaratne and Holford (1987, *Biometrics* **43,** 751–765) have given a fine explanation of the issues arising in cohort and case–control studies. But I was puzzled by its final sentence to the effect that the various criteria for identifying confounders could be stated loosely as “a variable *C* is a confounder if it is related to disease and also related to exposure.”

It was not clear to me just what that was supposed to mean—what behavior on the part of the statistician or the epidemiologist was being called for? Asking around, I learned that the practice called for in various epidemiology texts, and as followed by many epidemiologists, is that in a statistical analysis, adjustment should be made for a covariate—whether by stratification, matching, or possibly even regression procedures—only if that covariate is a confounder.

In Mantel and Haenszel (1959), no mention was made of confounders. Rather, at various places in that paper, it is indicated that covariates related to disease (no mention of exposure) should be adjusted for. Such adjustment would serve both to reduce bias and to increase precision. In a sense, any variable related to disease might be considered a confounder, regardless of whether it was also related to the particular exposure under study.

I had occasion in the recent past to be critical, though not in publication, of a paper dealing with the occurrence of some form of cancer because of its failure to take age into account. Another statistician took the defensive—both the case and control groups had about the same average age or about the same age distribution. So no allowance had to be made for age. Apparently, no counterargument could be compelling, as the statistician had learned from his professors and also from his texts, that adjustment would not be needed as the criteria for something being a confounder had not been met. Incidentally, the conditions for not requiring adjustment for age (or some other factor) can be loose—the average ages are the same, or the average ages do not differ significantly, or the age distributions do not differ significantly.

If there were any difference in the average age or age distributions, adjustment for age would remove any biasing effect of the difference. But suppose the study had been so contrived that there was an exact balancing of age in case and control curves. Would adjustment then be needed?

The answer is yes, according to the dictum in Mantel and Haenszel (1959). The variation in the factor particularly under study would be greater if various age strata were combined than would hold within the separate age strata. The Mantel–Haenszel stratification approach would both iron out any age distribution differences as well as reduce the variance, i.e., increase the precision, of the difference.

One reaction that I got to this kind of argument was that if age (or some other factor) were ignored, it would result only in lost power and greater variation. If the statistician is ready to pay a price in lost power, well, let him. However, my reply is that that can result in a biased analysis (not simply a bias in an estimate). Mantel and Patwary (1961) made the point that sometimes an analyst may be interested in bringing out that some effect was nonsignificant. The recipe for accomplishing this we suggested was to conduct a sufficiently small and sufficiently imprecise study. The failure to adjust for age or other real factors could give rise to just such a biased analysis. Actually, in situations I have been told by my client, when I wanted to make a more effective analysis, that the client thought we were supposed to bring out that the agent or exposure had no significant effect. My reply was that it was only when an effective analysis turned out negative that I could claim no effect—I should first do my best to bring out an effect.

But let me bring out a particular example [a more extreme illustration of which is given in Mantel and Haenszel (1959)]. Suppose that in one stratum we have a comparison of 10% vs 5%, for an odds ratio of 2.11. In another stratum the same odds ratio of 2.11 arises because of a comparison of 95% vs 90%. If the two strata were merged and all row totals were equal (such merger would be justified since the factor is apparently not a confounder by the usual definition), the comparison would be 52.5% vs 47.5%, for an odds ratio of 1.22.

In this example, the separate but equal odds ratios of 2.11 have merged to give a combined odds ratio of 1.22. Yet the separate but equal differences of 5% yield a combined difference also of 5%. The answer is that the percentages themselves and their differences are linear combinations of the observations but the odds and odds ratios are not. It was perhaps by mental analogy with other linear situations that some statisticians had thought that strata could be collapsed in certain situations. Even if precision were lost by collapsing, at least the estimate of effect would be preserved. But, as I have demonstrated, in nonlinear situations, the estimate of effect can become distorted.

But let me go back to the instance that had elicited my criticism, one in which age was ignored in a statistical analysis. Workers in the field of cancer know of the overwhelming influence of age in cancer incidence, which would make age adjustment essential in any epidemiologic study of cancer. Undoubtedly, age is highly important in many other morbidities and mortalities. However, age differs in an important way from other covariates for which we might wish to make adjustment.

Suppose we were making a cohort study in which individuals were followed in time. Then, in each successive year we would be studying that subsequent year's outcomes for individuals who had survived disease-free from the preceding year. A case–control study would not ordinarily give results comparable to those from a cohort study. Yet, by the simple expedient of matching or stratifying on age, with age strata as fine as feasible, the case–control or retrospective study can be taken to parallel a cohort study. Unfortunately, investigators frequently use overbroad age intervals, like 20 years, when stratifying on age. Given a typical threefold increase in cancer rates per 10-year increase in age, much narrower age strata would be essential.

Those statistician-epidemiologists to whom I have spoken have assured me that the practice of adjusting only for those "confounders" as defined is the rule generally followed. One cannot readily glean from any report that there has been anything amiss, for nearly every report gives assurance to its readers that things have been done just right. But it is likely that a large part of the epidemiologic literature has gone astray on this point.

Actually, statistical analysis would be facilitated by the rule of adjusting for variables related to disease, without concern as to whether they might not also be related to exposure status. Also, if there were several candidate exposures to which a study was directed, there would be no need to vary the factors on which stratification was to be made—the known or suspected factors influencing disease would remain the same whichever candidate exposure was being considered. But one or more other candidate exposures might be stratified on when considering each individual candidate exposure.

References

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22,** 719–748.

Mantel, N. and Patwary, K. M. (1961). Interval estimation of single parametric functions. In *Proceedings of the* 32*nd Session of the International Statistical Institute*, Tokyo. *Bulletin of the International Statistical Institute* **38,** 227–240.

Wickramaratne, P. J. and Holford, T. R. (1987). Confounding in epidemiologic studies: The adequacy of the control group as a measure of confounding. *Biometrics* **43,** 751–765.

Received July 1988.

RESPONSE

The interpretation of observational studies is a problem that statisticians have struggled with for some time now, and no doubt we will continue to grapple with these issues. In epidemiologic studies the probems are especially important because the questions being addressed deal with the physical and mental health of our society, and debate on the manner in which such studies should be designed and evaluated continues unabated (Feinstein, 1988). We are grateful for the provocative reactions to our work on confounding caused by the selection of a control group that have been contributed by Sander Greenland, Paul Holland, and Nathan Mantel. We hope that their reactions together with our response will help to shed further light on at least some aspects of these controversial issues.

To begin, we would like to thank Greenland for providing an updated bibliography on confounding and related topics. It should be clear that this is an area of research that continues to be active, and all of the papers he cites, with the exception of the papers by Rubin and Rosenbaum, appeared while our manuscript was in press, or in some cases after our paper was published. We also wish to clarify two points regarding his comments on the literature.

First, it is our understanding that the work of Rubin and Rosenbaum, cited by Greenland, is directed toward the problems of "defining, detecting, preventing and adjusting for" *bias in observational studies*, not the inadequacy of the control group, which they view as arising from "nonignorability of treatment assignment." In our opinion, the "nonignorability of treatment assignment" differs conceptually from the "inadequacy of the control group," a distinction that we have attempted to clarify in our response to Paul Holland's comments.

Second, Greenland states that "... confounding and comparability can be defined without any reference to covariates, a point not mentioned by Wickramaratne and Holford, but which follows directly from Rubin's formalization." We found this statement somewhat perplexing since all of our basic *definitions* of confounding and no confounding, both in the population and in the sample, were made *without any reference to covariates*. For example, no confounding in the population was defined as $P(D \mid \bar{E}) = P'(D \mid \bar{E})$ (§3.1 of our paper) and no confounding in the sample as $\mathrm{E}[\hat{P}(D \mid \bar{E})] = P'(D \mid \bar{E})$ (equation 4, p. 754 of our paper). The observant reader will note that there is *no covariate* involved in either of these definitions.

Mantel raises the important question often facing the applied statistician of what to do when faced with the analysis or design of a particular study. This is a question that was not specifically addressed in our paper, which dealt instead with establishing a framework to differentiate formally between the phenomena that can contribute to bias in epidemiologic studies. Reasons given by Mantel for covariate adjustment are "... to reduce bias and to increase precision." The particular example described by Mantel involves age as a potential confounder for cancer, a situation in which there is no question of whether there is in fact an association. However, in other situations, one must decide whether to adjust on an empirical basis, and in these instances it is not always obvious how one should behave. Statistical significance is not always the best guide as to which variables are confounders by any reasonable criterion, as was elegantly pointed out in an example given by Breslow and Day (1980, pp. 106–108). In this instance, the potential confounder was not significantly associated with disease, and yet the inference on the disease factor association was quite different depending on whether one controlled for the confounding variable in the analysis. Hence, we concur that with respect to bias, it is not always easy in practice to determine whether a variable should be controlled in the analysis.

In the analysis of covariance, where the linear model holds, the reduction in the standard error of the treatment effects realized by covariate adjustment is easy to see, even when the covariate is balanced by randomization. However, as pointed out by Mantel, the intuition gained from linear models can be misleading when it comes to the nonlinear estimators used in categorical data analysis. We do not agree with Mantel's second reason for covariate adjustment, in which he implies that precision necessarily increases when one adjusts for a covariate related to the response. This point was also discussed by Breslow and Day (1987) and by Day, Byar, and Green (1980). To illustrate, consider the hypothetical data shown in Table 1. In this case, there is an association between the confounder and disease, estimated by 1.622 (s.e. = .07313) for the regression coefficient in a linear logistic model, so that by Mantel's criterion this would qualify as a covariate that should be included in the analysis. The log odds ratio for the exposure disease association is .7293, whether one looks at the collapsed table or the stratified tables, so that there is no bias here when one does not adjust for the covariate. However, a stratified analysis using the linear logistic model gives a variance estimate of .05044, while the estimate from the collapsed table is slightly less, .04867. As this example illustrates, one can lose precision by unnecessarily adjusting for a covariate, which is just part of the reason for continued interest in this question.

Table 1
Hypothetical example of the distribution of subjects by level of a confounder, exposure to a factor, and disease status

Confounder	Factor	Not diseased	Diseased	Odds ratio
1	Not exposed	1,296	100	2.07
	Exposed	900	144	
2	Not exposed	2,304	900	2.07
	Exposed	1,600	1,296	
Total	Not exposed	3,600	1,000	2.07
	Exposed	2,500	1,440	

Finally, we wish to thank Holland for his thought-provoking comments on our paper. We are encouraged that our intuitive, *informal* description of confounding agrees in many ways with his own views. However, we do not agree with his statement that our "... *formal model* for confounding in the population is incomplete and unable to express important features of the problem that (we) *do* express informally in (our) discussion."

Although we welcome Holland's lucid exposition of Rubin's model and certainly do not dispute the fact that this model may shed further light on discussions of confounding, it is clear that it makes no real difference to the discussion of the specific issues that we have considered. Furthermore, our "new, nonstandard notation" was created with specific objectives in mind, which we shall attempt to clarify in the discussion that follows.

The quantity P' is a probability exactly like P; the reason we have used P' instead of P is to enable us to differentiate between proportions (or probabilities) in the *hypothetical* unexposed population (exposed group, if they were not exposed) and the *actual* unexposed population. We do not agree that P' has the "wrong thing in the conditioning." We are merely conditioning on a different variable than the one in Rubin's model. While Rubin's model conditions on the *observed* exposure status, we are conditioning on exposure status regardless of whether it is observed or hypothetical. In our notation, when the exposed group is *unexposed*, the appropriate exposure status (which is the variable on which we are conditioning) is $\bar{E}$ rather than E. Since we now have two populations (one hypothetical, the other actual) whose exposure status we consider to be $\bar{E}$, we used the prime notation to differentiate quantities in the hypothetical unexposed population from those in the actual unexposed population. With this notation, we can decompose the quantity $P'(D \mid \bar{E})$

in the usual manner as follows:

$$P'(D \mid \bar{E}) = \textstyle\sum_i P'(C_i \mid \bar{E})P'(D \mid C_i\bar{E}),$$

where $P'(C_i \mid \bar{E})$ is the proportion of individuals in the hypothetical unexposed population with the value of $C = C_i$ and $P'(D \mid C_i\bar{E})$ is the proportion of diseased individuals in the C_ith stratum of the hypothetical unexposed population. Note that C_i is now the value of C in the hypothetical unexposed population, and *not* the value of C in the exposed population. We agree with Holland that the wording of the definition of $P'(D \mid C_i\bar{E})$ in our paper is ambiguous. However, it is not at all clear to us that if $P'(D \mid C_i\bar{E})$ is interpreted as stated in Holland's equation (9), the assumption made in his equation (8) could be justified.

It seems to us that what disturbs Holland about our "unusual" notation is the fact that the "prime" ($'$) in $P'(\cdot)$ is outside the parentheses. We could easily shift it inside the parentheses by letting $\bar{E}'$ denote the hypothetical unexposed population, and defining $P(D \mid \bar{E}')$ to be the proportion of diseased individuals in this population. $P(C_i \mid \bar{E}')$ will be the proportion of individuals with a value of $C = C_i$ in this population and $P(D \mid C_i\bar{E}')$ will be the proportion of diseased individuals in the C_ith stratum of this population. This notation can now be considered "standard" by purists and, in addition, will give results identical to those discussed in our paper.

It would seem, at this point, that the only criticism of our model (by Holland) that we are left with is the fact that we have no apparatus for formally describing whether the value of the potential confounder C is affected by exposure. This deficiency can be easily remedied by introducing the notation $C(E)$, $C(\bar{E}')$, $C(\bar{E})$ to denote the values of C in each of the three populations—exposed, hypothetical unexposed, and actual unexposed. [This notation was introduced and discussed in Wickramaratne's unpublished Ph.D. thesis (Yale University, 1984), a discussion that was not included in our paper due to space constraints.] Clearly, if the value of C is not affected by exposure, then $C(E) = C(\bar{E}')$ for all individuals in the exposed population.

We agree that Rubin's model, because it is formulated at the unit level, brings an enviable explicitness and clarity to discussions of causal inference. What does not seem to be recognized is the fact that if our model was formulated at the unit level, it could be made just as explicit, but would *intentionally* have a rather different structure. In the paragraph that follows, we will outline briefly the manner in which *we* would formulate our model at the unit level.

Let $y_j(E)$ denote the disease status of the jth individual in the exposed group (where $j = 1, \ldots, N$) and let $y_j(\bar{E}')$ denote the disease status of this individual, if this individual had not been exposed. Let $y_k(\bar{E})$ denote the disease status of the kth individual in the unexposed group (where $k = 1, \ldots, M$). Letting E, $\bar{E}'$, and $\bar{E}$ index the exposed group, hypothetical unexposed group, and actual unexposed group, respectively, we now define the quantities $P(y = D \mid g = E)$, $P(y = D \mid g = \bar{E}')$, and $P(y = D \mid g = \bar{E})$ as the proportion of individuals with the disease in the groups indexed by E, $\bar{E}'$, and $\bar{E}$, respectively. We can then simplify this notation by letting $P(D \mid E)$, $P(D \mid \bar{E}')$, and $P(D \mid \bar{E})$ represent these quantities. Similarly, we can define $C_j(E)$, $C_j(\bar{E}')$, and $C_k(\bar{E})$ to represent the extraneous variable C, for individuals in each of these groups.

This notation was created in an attempt to emphasize the fact that the *population* of interest is the collection of individuals in the *exposed* group. Our interest lies in the collection of individuals in the *unexposed* group *only* insofar as it enables us to draw inferences about parameters in the hypothetical unexposed group, $\bar{E}'$. Consequently, we have no interest in, and hence no need for notation that allows us to represent the hypothetical disease status of an *unexposed* individual if that individual were exposed.

In our view, the fundamental difference between our model and Rubin's is not really a matter of notation or the fact that ours is at the subpopulation level, while his is at the unit level, but rather the manner in which the problem has been conceptualized. Rubin's model, which is based on the concept of one population assigned to two or more treatments, is rooted in the traditions of experimental design and randomization. It is our understanding that within this framework the model allows for the discussion and analysis of observational studies by allowing the assignment of individuals to treatment groups to be performed in a nonrandom manner.

Our notation, on the other hand, was created in part to reflect what we consider to be the essential asymmetry that exists between the treated (exposed) population and the control (unexposed) population in epidemiologic studies. As noted by Cochran (1983, pp. 42–44) and Miettinen (1985, pp. 30–34), among others, it is the characteristics of the treated population that dictate the appropriate choice of the control group; we think that our notation helps to illustrate this point well. Furthermore, it is our opinion that our model helps to provide a convenient framework for discussing and formalizing fundamental design issues in epidemiology, such as matching, the selection of an appropriate control group, and the use of multiple control groups. For these reasons, we believe there is room for more than one kind of notation in discussing problems of causality, especially when the different notation clarifies different aspects of the problem.

References

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research, Volume* I. *The Analysis of Case–Control Studies*. Lyon: IARC Scientific Publications.

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research, Volume* II. *The Design and Analysis of Cohort Studies*. Lyon: IARC Scientific Publications.

Cochran, W. G. (1983). *Planning and Analysis of Observational Studies*. New York: Wiley.

Day, N. E., Byar, D. P., and Green, S. B. (1980). Overadjustment in case–control studies. *American Journal of Epidemiology* **112,** 696–706.

Feinstein, A. R. (1988). Directionality in epidemiologic research. *Journal of Clinical Epidemiology* **41,** 705–707.

Miettinen, O. S. (1985). *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*. New York: Wiley.

Received July 1989; *revised September* 1989.

Priya J. Wickramaratne
Columbia University and
New York Psychiatric Institute
722 West 168th Street, Box 14
New York, New York 10032, U.S.A.

and

Theodore R. Holford
Yale University
New Haven, Connecticut 06510, U.S.A.